Integrated Access to Biological Data.
A use case

Marta González

Fundación ROBOTIKER,
Parque Tecnológico Edif 202
48970 Zamudio, Vizcaya – Spain
marta@robotiker.es

Abstract. This use case reflects the research on different and innovative ways to handle biological data repositories by means of semantic and artificial intelligence technologies such as ontologies, intelligent agents, semantic grid, etc. The human genome sequencing has given rise to a great number of biological data repositories that once analysed will be very essential for the study of diseases, pharmaceutical research, new treatments and for the development of new bio products. The problem faced is the huge quantity and heterogeneity of this kind of data and the also huge number and diversity of ontologies defined to model biological data.

1 Introduction

The aim of this use case is to provide an unified access point to diverse biological data repositories: accessible through internet (Nucleotide Sequences, amino acid sequences,…), corporate databases, results of experiments (DNA-chips), health cards, medical literature sites…This unified access has to be provide with the purpose of generation and extraction of knowledge from biological data by means of ontologies, combining them (ontology merging) and/or associating them (ontology mapping) to be exploited by means of annotations, intelligent agents, semantic web services and/or semantic grid.

Currently, a great diversity of biological data repositories exists: databases accessible through Internet, corporate databases and microarrays experiments results among others. Equally exists a great diversity of ontologies to model this data. Therefore the situation the researchers has to face with is a lot of disperse data and different disconnected and poor friendly tools to access such data, therefore the researches have to confront great difficulties to aggregate all the data to carry out the research tasks in an integrated way.

Up to now ontologies in biology were considered as mere guides for data structure, with the only purpose to access to the more adequate documents and articles to the researcher interests. This new vision will allow, combining and associating existing ontologies in the biological field, an integrated modelling of the biological data sources (genomics, proteomics, metabolomics and systems biology).

Once modelled, the annotations, intelligent agents, semantic web agents and the semantic grid will offer a centralised access point to extract and generate knowledge from the biological data repositories.

2 Biological Data Inherent Features

The inherent features of the biology field are: huge quantity of disperse-distributed-and autonomous data with great difficulties to be integrated due to their differences in: terminology (synonyms, aliases, …), syntax (spelling, file structure, …) and semantics (intra-/interdisciplinary homonyms)[1]. As instance it is complicated to know if a database table called “Species” is the same table called “Organisms” in other different database.

The following figure (Fig 1) illustrates the semi-exponential growth of DNA databases along the years:

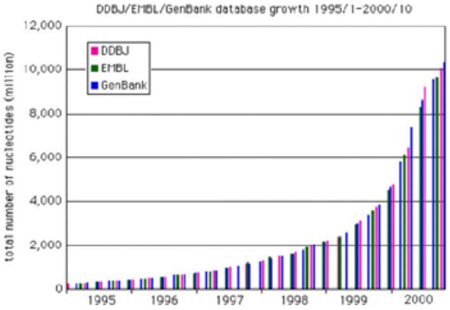

Fig. 1 Semi-exponential growth of DNA databases along the years (Source [2])

In order to highlight the inherent biology features it will be cited the most important biological data repositories and a short state of the art related to semantic technologies that could be applied to the biological domain.

2.1 Biological Data Repositories

Currently, more than 500 biological data repositories exist (584 according to "The Molecular Biology Database Collection: 2004 update”), the following repositories are considered the main or reference ones:

2.1.1 Nucleotides Sequences

These three databases are synchronized and daily updated.

- EMBL (United Kingdom)Nucleotide Sequence Database: it is the main European nucleotide sequences repository (http://www.ebi.ac.uk/embl/).
- GenBank (USA): GenBank is the National Center for Biotechnology Information (NCBI) genetics sequences(http://www.ncbi.nlm.nih.gov/Genbank/index.html).

- DDBJ (Japan): it is the unique DNA database in Japan, officially certified to gather DNA sequences for researches. (http://www.ddbj.nig.ac.jp).

2.1.2 Amino acid sequences

- SwissProt: protein sequences database (http://us.expasy.org/sprot/).
- PIR (Protein Information Resource): protein sequences database (http://pir.georgetown.edu/).
- PDB (Protein Data Bank): repository for the processing and distribution of 3-D biological structure.(http://www.rcsb.org/pdb/).

2.1.3 Gene Expression

- GDX , once of the first databases that integrates diverse gene expression data types and that was developed before biochips irruption.
- ExpressDB relational database containing yeast and E. coli RNA expression data. (http://arep.med.harvard.edu/ExpressDB).

2.1.4 Scientific literature

Many organisms offer scientific literature freely or by subscription. **MEDLINE** is health information from the world's largest medical library, the National Library of Medicine.

PubMed[8], a service of the National Library of Medicine, includes over 15 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

UpToDate[9] Specifically designed to answer the clinical questions that arise in daily practice and to do so quickly and easily so that it can be used right at the point of care. The Topic Reviews are written exclusively for UpToDate by physicians for physicians. The content is comprehensive, yet concise, and is fully referenced. It goes through an extensive peer review process to insure that the information and recommendations are accurate and reliable.

2.1.5 Corporate Databases

Companies owns corporate databases with their research labours results stored, as instance, biochips experiments results. At the current situation the results of biochips experiments are stored at private researchers' databases. The boost that these techniques are winning in the biomedical research domain, along with their use extension, is helping as an important engine for the development of public databases with data from experiments; where these data can be stored for later analysis and comparison.

2.1.6 Health Cards

The Health Cards appear as a future possibility to store data that can be computer read and that it is issued to patients or sanitary professionals to facilitate the medical care attention. To store data in a card that can be computer accessible various technologies exist: magnetic strip, integrated circuit memory cards and optical memory cards. The utility of these cards could be administrative tasks, emergency health cards, specific care records and patients general medical records.

The information to be stored is under discussion, in Europe some efforts are focused on patients mobility comfort. This comfort it is desired to be reached by forgetting paper forms and adopting health cards. This means an unification of data to be stored, the media to be used and medical nomenclature. Efforts as SNOMED or GALEN are faced to this last objective.

Another use of the health cards could be the possibility to match genetic patient data with biological databases to know as instance the probability of a patient to suffer a certain disease. Also this could be applied to patients communities for statistical research in order to define priorities and strategies for health and social security and policy, planning and evaluation of preventive measures and care services and cost-benefit calculations between preventive measures and therapeutic actions.

3 Use of ontologies

Currently a huge quantity of ontologies, in the biological domain, have been defined along with specific purpose data mining tools. The data mining tools, that elaborate analysis models, as instance gather genes or experiments in function of different patterns, solve the immediate problems that the researchers have to face with, while ontologies model knowledge for the future research assistance. The problem lie in ontologies definition phase, in definition or not defined at all. In some cases a browser allows to look for information at the repositories by using an ontology[1]. Some initiatives are focused on prospective analysis of diverse biological repositories interaction, as instance: BIOINFOMED, BIRN network, e-BioSci, HKIS, INFOGENE, PRIDEH-GEN[3].

Text mining and natural language understanding in biology can also profit from ontologies. Where currently mostly statistical and proximity approaches are applied to text analysis ontologies can support parsing and disambiguating sentences by constraining grammatically compatible concepts.

To eliminate semantic confusion in molecular biology, it will be therefore necessary to have a list of the most important and frequently used concepts coherently defined so that e.g. database managers, curators and annotators could use such set of definitions either to create new software and database schemas, to provide an exact, semantic specification of the concepts used in an existing schema and to curate and annotate existing database entries consistently.

Efforts such as SNOMED provide us with vocabularies of 36,000 medical definitions and concepts, and ICD10 is the International Classification of Diseases. This kind of efforts allows to own a common language for interoperability, facilitating scientific research labours.

The following ontologies can be considered as a subset of the most widely known ontologies in the biological field:

Gene Ontology (GO) [4]Gene ontology is a controlled vocabulary that has been developed into the project OBO. GO describes how the gene products behave in a cellular context. Currently three ontologies are published at Internet: Biological

Process, Molecular Function and Cellular Component representing biological targets that a gene product contributes, biochemical activities of gene products and places where the gene products can be active. Currently some databases are annotated with GO terms.

The Microarray Gene Expression Data (**MGED**) [5] Society is an international organisation of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments. The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well annotated data within the life sciences community. The available ontology is in OWL format, it is composed by standard terms for the annotation of microarray experiments. These terms will enable structured queries of elements of the experiments. Furthermore, the terms will also enable unambiguous descriptors of how the experiments were performed.

The purpose of NLM's Unified Medical Language System® (**UMLS®**) [6] is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs) for use by system developers in building or enhancing electronic information systems that create, process, retrieve, integrate, and/or aggregate biomedical and health data and information, as well as in informatics research.

3.1 Ontology merging and mapping

Once the main ontologies are identified we need some mechanisms to join them in order to access to diverse data repositories by means of a unique ontology.

Ontology **Merging** allows to know which concepts in an ontology A are the same than in other ontology B. Detecting common concepts and allowing the “jump” between ontologies. The ontology merging could be used by a company that wants to use de facto standard ontologies -GO, MGED,...- associating them to specific company’s ontologies. This way proprietary repositories, as instance repositories with experiments results, can be “linked” with public ones. Also ontology merging allows to merge the repositories described by such ontologies (once repositories are annotated).

Ontology **Mapping** allows to sum up one ontology C with other ontology D to obtain a more complete ontology. Due to the fact that a protein could be implied in cell signalling as in a biological process, summing up two ontologies, one describing cell signalling and other describing biological processes can give us a general overview of a protein function.

3.2 Database annotation

Once we have been able to merge or map the ontologies, as described above, we will have to be capable of linking these resulting ontologies with public or proprietary data repositories through semantic annotation. Deep annotation[7] is a framework to be taken into account at this stage. Deep annotation is a framework to provide semantic annotation of large sets of data. It is used to describe the process that allows to derive mappings between information structures using information proper, information structures and information context.

Annotation is relevant for scientific databases, because scientific databases have been developed with the researchers community in mind, trying to stimulate cooperation. Some of the most known databases annotated are: GenBank, H-Invitational, Invitrogen, InterDom, KEGG and USCSC Genome Bioinformatics. GeneOntology has been used to annotate a great number of different databases, as instance: SGD, FlyBase, GO Annotations@EBI Arabidopsis, GO Annotations@EBI Human, GO Annotations@EBI Mouse, GO Annotations@EBI PDB, GO Annotations@EBI Rat, GO Annotations@EBI UniProt, GO Annotations@EBI Zibrafish,...

4 Conclusions

As semantic web technologies are still developing, also the tools needed to implement such technologies are also in a developmental stage, limiting the application of such technologies to the biological domain. Future advances in semantic web technologies could be applied to solve the immediate Scientifics needs: data aggregation and interoperability, unique entry point for data and processes, agreement in terminology, syntax and semantics related to biological data, semantic data annotation to turn human-understable data into machine-understable data, inference languages to extract and generate knowledge from aggregated data,…

References

1. RZPD Deutsches Ressourcenzentrum für Genomforschung GmbH http://www.bioinfo.de/isb/2002/02/0017/main.html#img-1
2. School of Mathematical Sciences (Israel) http://www.math.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec05/node3.html
3. Infogenmed Project Web Site. http://infogenmed.ieeta.pt/webdata/related-projs.html
4. Gene Ontology http://obo.sourceforge.net/main.html
5. MGED Ontologu http://mged.sourceforge.net/Ontologies.shtml
6. Unified Medical Language System – UMLS (http://www.nlm.nih.gov/research/umls/)
7. On Deep Annotation. S.Handschuh, S.Staab, R. Volz, WWW2002, University of Karlsruhe,2003.
8. PubMed Website, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
9. UpToDate WebSite http://www.uptodate.com/service/index.asp